WWWを用いた評価極性タグ付きコーパスの自動構築

鍛治伸裕　　喜連川優

東京大学 生産技術研究所
{kaji, kitsure}@tkl.iis.u-tokyo.ac.jp

1　はじめに

近年，ブログや掲示板の普及に伴い，インターネットには企業やその製品の評判が頻繁に書き込まれるようになった．そして，こうした情報を有効に活用するための手段として，評判情報を対象としたテキスト処理が注目されている [1, 2].

評判情報を処理するためには，単語や文の評価極性 (肯定的か否定的か) の扱いが重要となる．そして，そのために利用されるのが，語句の評価極性を登録した辞書 [3] や，評価極性がタグ付けされたコーパス (評価極性タグ付きコーパス) などである．

本論文では，この評価極性タグ付きコーパスの構築方法を議論する．現在のところ，こうしたコーパスを構築する方法は 2 種類ある．1 つは人手で作成するという方法であり，もう 1 つはインターネットのレビューサイトを利用する方法である．レビューサイトの投稿には，それが肯定的評価 (好評) なのか否定的評価 (不評) なのかを示すタグが投稿者によって与えられている場合があるので，そのまま極性タグ付きコーパスとして利用できる．

しかし，こうした構築方法には欠点がある．最も深刻なのは規模の問題である．大規模なコーパスを作ることを考えると，人手による方法はコストが高く現実的ではない．同様のことは，レビューサイトを利用する方法にも言える．なぜなら，レビューサイトから作られたコーパスにはしばしばノイズが混入するため，人手で修正を加えることが多く，そのコストが無視できないからである．さらに，レビューサイトに登録されているテキストの総数が，構築できるコーパスの大きさの上限になってしまうという問題もある．

こうした問題を踏まえて，任意の HTML 文書集合から，評価極性タグ付きコーパスを自動構築する方法を考案した．この手法は「箇条書き」「表」「定型文」といった記述形式を利用することにより，評価を表している文 (評価文) とその極性を HTML 文書から高い精度で抽出できる．実験のため提案手法を約 9,500 万件の HTML 文書に適用したところ，100,153 の評価文を獲得することができた．また，このうち 500 文を人手で調べたところ，90%を越える文に適切な極性タグが付与されていることが分かった．

2　コーパスの概要

まず，我々が自動構築する評価極性タグ付きコーパスについて簡単に説明をする．表 1 に，実際に自動構築されたコーパスの一部を示す．

構築されるコーパスでは，好評と不評の 2 種類の評価極性タグが使われる．表中ではそれぞれ＋と－で表されている．そして，評価極性タグは文単位で付与される．このとき 1 つの文に対して，2 つのタグを同時に付与するようなことはしない．また，評価極性を持たない文はコーパスに登録しない．

ここで注意を促しておきたいのは，文書単位などの複数文単位でタグを付与したコーパスを構築するのではない，ということである．このように決めた理由は，一つの文書には好評文と不評文が混在している場合があり，そのような文書に極性タグを付与しても活用するのが難しいと考えたからである．なお，レビューサイトを利用して構築された極性タグ付きコーパスは，複数文単位でタグが付与されたものが多い．このことも，従来のコーパス構築方法の問題点の一つである．

表 1: 自動構築された評価極性タグ付きコーパス (一部)

評価極性タグ	評価文
＋	順応性が素晴らしくある．
－	費用が高い．
－	いい加減な意見，ふざけた意見などが出てくる．
－	エンジンが非力で少々うるさい．
＋	使い方がわかりやすい．
＋	何と言っても，料金が良心的だ．

3　アイデア

では次に，このようなコーパスを自動構築する方法を具体的に説明していく．

我々のアイデアは「箇条書き」「表」「定型文」といった記述形式を利用して，HTML 文書から評価文とその極性を抽出するというものである．まず本節では，これら 3 つの形式で記述された評価文の例を概観して，基本的なアイデアを説明する．そして，次節でコーパスを自動構築する詳細な手順を説明する．

3.1　箇条書き

まず我々が着目したのは，図 1 のように箇条書き形式で列挙された評価文である．この箇条書きには「良い点」「悪い点」という見出しがついているので，これを利用すれば，そこに評価文が記述されていることが分かる．

以下では「良い点」「悪い点」などのように，評価文の存在を示す見出し表現を手がかり句と呼ぶ．特に肯定的な評価文 (好評文) の存在を示す手がかり句を「好評手がかり句」と呼ぶ．例えば「良い点」などである．同様に，否定的な評価文 (不評文) の存在を示す手がかり句を「不評手がかり句」と呼ぶ．

このような箇条書きから評価文を抽出するのが 1 つめの方法である．HTML 文書中の見出しと箇条書は，<h1> や <ol> などのタグを利用すれば判別可能である．また，手がかり句は，あらかじめ手がかり句のリストを作成しておくことにより認識できる．

良い点
- 変に加工しない素直な音を出す.
- 曲の検索が簡単にできる.

悪い点
- リモコンに液晶表示がない.
- ボディに傷や指紋がつきやすい.

図 1: 箇条書き形式で記述された評価文

3.2　表

次は，図 2 に示すような表形式である．この表は左側の列が見出しの働きをしているが，ここにも手がかり句 (気に入った点，イヤな点) が使われているので，表中に評価文があることが分かる．

燃費 (市街地)	7.0km/litter
燃費 (高速)	9.0km/litter
満足度	95%
気に入った点	4ドアなのにカッコよすぎる.
イヤな点	シートがぼろくライトが暗い，色がはげてきてる.

図 2: 表形式で記述された評価文

3.3　定型文

3 番目のアイデアは定型文を利用するというものである．次のような例文を考える．

(1)　これの良いところは計算が速いことです.
(2)　悪い点は，慣れるまで時間がかかること.

いずれの文も「良いところ/悪い点は〜なこと」という型を使って記述されている．ここでも手がかり句が使われていることに注目されたい．

こうした定型文はパターンを使って認識し，そして「計算が速い」「慣れるまで時間がかかる」という部分だけを評価文としてコーパスに登録することにした．文全体を登録するのではないことに注意されたい．

4　評価極性タグ付きコーパスの構築

次にコーパス構築の具体的な手続きについて述べる．まず準備として手がかり句リストを作成した．内訳は，好評手がかり句が 303 表現，不評手がかり句が 433 表現である．リストの一部を示す．

良い点，善い点，利点，メリット
悪い点，改善してほしい所, 難点，デメリット

そして，このリストを用いて，以下の手順でコーパスを構築する．(1) HTML 文書に前処理を施す．一部の箇条書きや見出しは HTML タグを使わないで記述されているので，ルールでタグを補完する．そして，その後にテキストを文単位に分割する．(2) 前節で説明した「箇条書き」「表」「定型文」から評価文とその極性を抽出する．(3) 抽出された評価文に後処理を加える．

本節では「箇条書き」「表」「定型文」から評価文を抽出する方法を順に説明し，最後に後処理を説明する．

4.1　箇条書きからの抽出

まず，箇条書きからの評価文抽出であるが，これは手がかり句リストと HTML タグを利用すれば容易に

実現できる．すなわち，手がかり句が見出しになっている箇条書きを見つけて，その箇条書きの項目を順に取り出していけばよい．例えば図 1 からは「変に加工しない素直な音を出す」「曲の検索が簡単にでる」が好評文として,「リモコンに液晶表示がない」「ボディに傷や指紋がつきやすい」が不評文として取り出される．

問題となるのは，1 つの項目に複数文が記述されている場合の処理である (図 3 の 3 番目の項目)．このような場合は，1 項目に好評文と不評文が混在している可能性があるため，抽出に使わないことにした．すなわち，図 3 からは「発色がものすごくよい」と「撮っていくうちに楽しくなる」だけが抽出される．

よい点

- 発色がものすごくよい．
- 撮っていくうちに楽しくなる．
- 400 万画素という高画素．200 万画素では物足りなかった．

図 3: 1 項目に複数文が記述されている箇条書き

4.2 表からの抽出

次に，表から評価文を抽出する方法を述べる．この処理で問題となるのは，HTML 文書中の表は多種多様な使われかたをするので，あらゆる表に対応した抽出規則を作成することは容易でないということである．

そこで，図 4 のような 2 つのタイプの表だけを考えることにした．タイプ A は，1 列目に手がかり句があって，その横に評価文があるタイプである．前節で紹介した図 2 はこのタイプに相当する．タイプ B は，1 行目に手がかり句があって，その下に評価文があるタイプである．図中の C_+ と C_- は好評手がかり句と不評手がかり句を表し，＋と－は好評文と不評文を表す．

タイプ A

C_+	＋	＋	＋
C_-	－	－	－

タイプ B

	C_+	C_-	
	＋	－	
	＋	－	
	＋	－	

図 4: 評価文を抽出するときに考慮する表タイプ

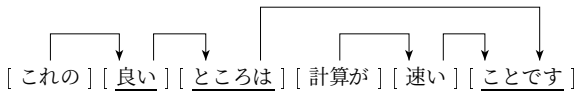
与えられた表のタイプは，1 列目 (1 行目) を調べて，好評手がかり句と不評手がかり句の両方が出現していればタイプ A(タイプ B) であると判定する．

表のタイプが決まれば，あとは図の＋と－に対応するマスから評価文を抽出すればよい．このときに留意すべきは次の 2 点である．まず，1 つのマスに複数文が記述されている場合は抽出対象としない．これは箇条書きの 1 項目に複数文が記述されている場合と同様の理由からである．また，詳細は省略するが，マスの中に箇条書きがある場合は，箇条書きから評価文を抽出するときと同様に処理する．

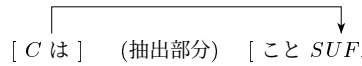
4.3 定型文からの抽出

定型文はパターンを使って認識し，そこから評価文を抽出する．この抽出方法の特徴は，依存構造を考慮したパターンを使う点である．

まず 3 節の例文を依存構造木に変換したもの考える．



下線部が「型」に相当する．抽出すべき表現 (計算が速い) は，下線を引いた文節の間に存在する．この型に「良いところ」という手がかり句が含まれることに着目して，以下のような抽出パターンを考案した．



C は手がかり句を表す，$SUFFIX*$ は接尾辞，判定詞，助詞，助動詞が 0 個以上あることを表す．$SUFFIX*$ は，上の例文では文末の文節「ことです」の「です」にマッチする．(抽出部分) は実際に評価文として抽出される部分である．上の例文からは「計算が速い」が抽出される．

4.4 後処理

最後に，抽出された評価文に後処理を加える．後処理が必要になるのは次のような場合である (図 5)．

燃費 (市街地)	10.0km/litter
燃費 (高速)	12.0km/litter
気に入った点	スタイリング
イヤな点	テールランプ周りの造形

図 5: 評価対象しか書かれていない表

この表からは「スタイリング」と「テールランプ周りの造形」が抽出されてしまう．これは，表に評価対象しか書かれていないのが原因であるが，このような文がコーパスに登録されるのは望ましくない．

そこで，こうした文を後処理で削除することにした．評価対象は名詞句であることがほとんどなので，抽出された評価文を解析して，それが名詞句であればコーパスに登録しないようにした．

5 実験と考察

コーパスの自動構築実験 約 9,500 万件の HTML 文書を用いて，評価極性タグ付きコーパスの自動構築実験を行った．その結果，100,153 の評価文を抽出することができた．好評文は 51,663 文，不評文は 48,490 文であった (抽出された評価文の具体例は表 1 を参照)．

人手による評価と考察 自動学習されたコーパスから 500 文を無作為に抽出して，文に妥当な極性タグが付与されているかを二人の被験者 (被験者 A，B と呼ぶ) が調べた．

被験者には，上記の 500 文に対して「好評」「不評」「不定・極性無し」のいずれかのタグを割り当てるよう指示した．「不定・極性無し」は，好評とも不評ともとれる文や，評価極性をもたないと考えられる文に付与される．ただし，この作業中，自動付与されたタグや元の HTML 文書は被験者に提示していない．

この実験で，二人の被験者が同一のタグを付与したのは 500 文中 467 文であった．そして，コーパスに付与されているタグと被験者 A が付与したタグが一致したのは 500 文中 459 文であった．ただし，被験者が「不定・極性無し」と判断した文は不一致に数えている．同様に被験者 B の場合は 460 文であった．被験者間のタグが一致した文が 467 文であることを考えると，自動構築されたコーパスには，非常に高い精度で適切なタグが付与されていると結論づけることができる．

コーパスのタグと被験者のタグが一致しなかった文を観察した結果，そのほとんどは，評価極性が強く文脈に依存する文であった．例えば，コーパスには「何しろ情報量が多い」が好評文として登録されていたが，被験者は二人とも「不定・極性無し」と判断していた．

分類タスクへの適用 このコーパスを評判情報処理に利用したときの有効性を検証した．ここではタスクの一例として評価文を好評と不評に分類するタスクを取り上げる．提案コーパスをトレーニングデータに使って分類器を学習し，評価文の分類実験を行った (表 2)．

分類器は単語を素性とするナイーブベイズを採用した [2]．テストデータはインターネットの掲示板から取得した．コンピュータに関するものが 2 つ (異なる掲示板から取得した)，レストランと自動車に関するものが 1 つずつである．なお，テストデータと自動構築されたコーパスには重複がないことを確認している．

分類実験の結果，全てのテストデータに対して 80% を越える高い精度が観察できた．この結果から，提案コーパスは実タスクでも十分利用できる質であると考えている．なお，分類精度は，テストデータのドメイン間で顕著な差は見られなかった．

表 2: 分類実験の結果

	文数	(好評文/不評文)	分類精度
コンピュータ (1)	681	(329/362)	0.810
コンピュータ (2)	1,162	(614/548)	0.832
レストラン	1,162	(753/409)	0.807
自動車	1,856	(1,056/800)	0.821

6 おわりに

本論文では，評価極性タグ付きコーパス構築を自動構築する手法を提案した．そして，実験を通して，提案手法が大規模で実用的なコーパスを構築可能であることを示した．

今後の課題としては，より大規模な HTML 文書からコーパス構築を行うことや，評価文抽出規則を拡張することを考えている．また，構築したコーパスを利用して，依存構造など深い情報を利用した評価文の分類，抽出手法を考案していきたい．

参考文献

[1] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. Mining product reputations on the web. In *Proceedings of the KDD*, 2002.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the EMNLP*, 2002.

[3] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientation of words using spin model. In *Proceedings of the ACL*, pp. 133–140, 2005.